OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM

Yutao Hu[1,2]*, Tianbin Li[2]*, Quanfeng Lu[2]*, Wenqi Shao[2]†,
Junjun He[2], Yu Qiao[2], Ping Luo[1,2]†
[1]The University of Hong Kong [2]Shanghai AI Laboratory

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in various multimodal tasks. However, their potential in the medical domain remains largely unexplored. A significant challenge arises from the scarcity of diverse medical images spanning various modalities and anatomical regions, which is essential in real-world medical applications. To solve this problem, in this paper, we introduce OmniMedVQA, a novel comprehensive medical Visual Question Answering (VQA) benchmark. This benchmark is collected from 73 different medical datasets, including 12 different modalities and covering more than 20 distinct anatomical regions. Importantly, all images in this benchmark are sourced from authentic medical scenarios, ensuring alignment with the requirements of the medical field and suitability for evaluating LVLMs. Through our extensive experiments, we have found that existing LVLMs struggle to address these medical VQA problems effectively. Moreover, what surprises us is that medical-specialized LVLMs even exhibit inferior performance to those general-domain models, calling for a more versatile and robust LVLM in the biomedical field. The evaluation results not only reveal the current limitations of LVLM in understanding real medical images but also highlight our dataset's significance. Our code with dataset are available at https://github.com/OpenGVLab/Multi-Modality-Arena.

1. Introduction

Recently, Large Vision-Language Models (LVLMs) have exhibited remarkable advancements across various domains, including embodied AI [85], autonomous driving [121], and remote sensing [67]. Encouraged by their achievements, a growing number of LVLMs tailored for medical applications have emerged, claiming impressive performance across a wide spectrum of medical challenges [71, 84, 117, 127]. However, despite the growing attention these models have obtained, there has been a noticeable lack of comprehensive evaluations, particularly when it comes to real medical images, which strongly hinders a thorough understanding of their applicability and performance in medical contexts [116].

Table 1. The comparison of the number of modalities, images and question-answering items in different medical VQA datasets. † indicates we calculate the numbers by ourselves, without the official statistic could be directly adopted.

Dataset	# Modalities	# Images	# QA Items
VQA-RAD [69]	3	315	3515
SLAKE [75]	3	642	14,028
Path-VQA [122]	2†	4998	32,799
VQA-Med [24]	5†	4500	5500
OmniMedVQA	12	118,010	127,995

We attribute this challenge to the absence of a comprehensive and diverse evaluation benchmark, one that encompasses images captured from various modalities and covers a broad spectrum of human anatomies. In more detail, the ability to answer questions based on a given image is fundamental, yet critically important in evaluating the performance of LVLMs. To facilitate this purpose, a comprehensive Visual Question Answering (VQA) dataset is indispensable. However, as indicated in Table 1, the majority of existing VQA datasets suffer from size limitations. Moreover, many of them provide only a limited number of modalities and focus exclusively on specific aspects of human anatomy. Consequently, these datasets do not meet the requirements for a comprehensive evaluation of LVLMs in the medical domain.

To address this challenge, this paper introduces OmniMedVQA, a large-scale and comprehensive Visual Question Answering benchmark designed for the medical domain. Considering the scarcity of medical image-text data, we collect numerous medical classification datasets and then transfer these data to VQA format according to their

*Equal contribution.
†Corresponding author.

Table 2. **Comparison of Different LVLMs.** VE, ToP and TuP indicate the visual encoder, number of total parameters and tuning parameters, respectively. † indicates that the model is frozen. CC* consists of COCO [34], CC3M [105], and CC12M [32]. CC, VG, SBU CY, and L400 indicate Conceptual Caption [32, 105], Visual Genome [66], COYO-700M [27] and LAION 400M [101], respectively. LLaVA-I and G4L represent 158K multimodal instruction-following data in LLaVA [77] and data generated by GPT-4 for building an instruction-following LLMs [92]. QA* denotes 13 question-answering datasets in InstructBLIP [41].

Model	Model Configuration				Image-Text Data		Visual Instruction Data	
	VE	LLM	ToP	TuP	Source	Size	Source	Size
BLIP2 [73]	ViT-g/14†	FlanT5-XL†	4B	107M	CC*-VG-SBU-L400	129M	-	-
LLaVA [77]	ViT-L/14†	Vicuna	7B	7B	CC3M	595K	LLaVA-I	158K
LLaMA_Adapter_v2 [48]	ViT-L/14†	LLaMA†	7B	63.1M	L400	200M	LLaVA-I+G4L	210K
MiniGPT-4 [128]	BLIP2-VE†	Vicuna†	7B	3.1M	CC-SBU-L400	5M	CC+ChatGPT	3.5K
mPLUG-Owl [123]	ViT-L/14	LLaMA†	7B	1.1B	CC*-CY-L400	204M	LLaVA-I	158K
Otter [70]	ViT-L/14†	LLaMA†	9B	1.3B	-	-	LLaVA-I	158K
InstructBLIP [41]	ViT-g/14†	Vicuna†	7B	107M	-	-	QA*	16M
VPGTrans [125]	ViT-g/14†	Vicuna†	7B	107M	COCO-VG-SBU	13.8M	CC+ChatGPT	3.5K
Med-Flamingo [84]	ViT-L/14†	LLaMA †	8.3B	1.3B	MTB, PMC-OA	2.1M	-	-
RadFM [117]	ViT-3D	LLaMA	14B	14B	MedMD	16M	RadMD	3M
MedVInT_TE [127]	ResNet-50	LLaMA†	7B	156.4M	PMC-OA	1.64M	PMC-VQA	152k
LLaVA-Med [71]	ViT-L/14†	Vicuna	7B	7B	PMC-15M	600K	PMC-15M + GPT4	60K

classification attribute based on the powerful context reasoning capacity of GPT. Generally speaking, OmniMedVQA boasts two primary highlights. First, it encompasses images from 12 different modalities, including MRI, CT, X-Ray, histopathology, fundus photography, *et al.*, resulting in a highly diverse dataset. Importantly, all these images originate from real medical scenarios, aligning OmniMedVQA closely with real-world applications. Second, OmniMedVQA covers over 20 distinct human anatomical regions. As illustrated in Fig 1, OmniMedVQA spans from the brain to the extremities, which facilitates a more comprehensive evaluation of different LVLMs and calls for a more versatile medical LVLM. Moreover, for the convenience of evaluation, we assign the incorrect options to each question-answering (QA) pair, transferring our OmniMedVQA to a multi-choice Question-Answer dataset. Overall, our OmniMedVQA dataset contains 118,010 different images with 127,995 different test items, leading to a large-scale evaluation benchmark.

In our evaluation, we assess a total of twelve representative models, including eight general-domain LVLMs, *e.g.,* BILP2 [73], MiniGPT-4 [128], InstructBLIP [41], mPLUG-Owl [123], Otter [70], LLaVA [77], LLama_adapter_v2 [48], and VPGTrans [125], as well as four specialized medical LVLMs, including Med-Flamingo [84], RadFM [117], MedVInT [127], and LLaVA-Med [71]. Notably, since OmniMedVQA is extremely challenging, especially for the general-domain LVLM, we find it is difficult for the model to directly generate the answer even if we give them the candidate options. To better evaluate their inherent knowledge in the biomedical domain, we adopt two different metrics, Question-answering score and Prefix-based Score [73, 120] to calculate the VQA accuracy, leading to a more comprehensive evaluation. Through our extensive experiments, we surprisingly find, that medical-specialized LVLMs exhibit superior performance compared to general-domain LVLMs. Specifically, although medical LVLMs obtain better performance on some specific modalities such as CT, MRI and X-Ray, they struggle to consistently outperform general models across all modalities, particularly those with similar distributions to general images. Furthermore, we emphasize the pressing need for a robust model that can effectively align image-text pairs in the medical field. Such a model is crucial for medical-domain LVLMs, as it can generate accurate and comprehensive descriptions for medical images and support the sufficient training of LVLMs.

We want to emphasize that, although the classification attribute is compact, it does provide a basic evaluation benchmark for the medical area and mitigate the collecting cost. More importantly, according to the evaluation results, although not complex, the existing LVLMs, especially those medical-specialized LVLMs, do not exhibit satisfactory performance on our OmniMedVQA dataset, which not only shows the shortcoming of these models but also demonstrates the challenge of the proposed dataset.

The main contributions of this paper are summarized as follows:

- We propose OmniMedVQA, a large-scale and comprehensive Visual Question Answering benchmark tailored to the medical domain. OmniMedVQA contains 12 different modalities and covers more than 20 unique human anatomical regions, establishing a comprehensive benchmark for evaluating the fundamental capabilities of LVLMs in addressing medical challenges.

- We conduct a thorough evaluation for 12 different LVLMs, including 8 general-domain LVLMs and 4 specialized LVLMs designed for medical applications. As far as we know, it is currently the most comprehensive evaluation of LVLMs towards the medical domain.
- Our evaluation uncovers several innovative insights and provides valuable guidance for improving LVLMs toward medical applications in the future.

2. Related Work

2.1. Large Vision-Language Models

Based on the recent emergence of Large Language Models (LLMs) such as LLaMA [111] and GPT [90], LVLMs utilize the knowledge from LLMs and align visual features to the textual space for various text output. Flamingo [18] is one of the early attempts that insert cross-attention layers into LLMs to introduce visual features into textual space. Meanwhile, to better align multi-modal features, BLIP2 [73] unifies the pre-trained visual encoder with LLM through an ingeniously designed Q-former. After that, InstructBLIP [41] extends BLIP-2 with instruction-following data and obtains better performance. Motivated by this success, most LVLMs are built through the instruction-tuning pipeline. For example, LLaVA [77] constructs 158K instruction-following data to conduct the training process and achieves great performance. Building upon the success of LLaVA, several subsequent LVLMs [48, 70, 123] leverage the high-quality 158k multimodal data to facilitate the training process. Furthermore, MiniGPT-4 [128] aligns a frozen visual encoder with a frozen LLM via only one projection layer. To better fine-tune the model, MiniGPT-4 utilizes 3500 detailed image-description pairs, illustrating that even a relatively small amount of high-quality data can significantly enhance the training of LVLMs. Additionally, VPGTrans [125] transfers the text encoder of BLIP2 model to Vicuna, which reduces the training costs and maintains the convincing performance.

Recently, encouraged by the success of these general-domain LVLMs, researchers have embarked on the development of LVLMs for the medical field. Med-Flamingo [84] is the pioneering effort in this field, which extends the Flamingo into the medical domain by pre-training on multi-modal knowledge sources across medical disciplines. Meanwhile, LLaVA-Med [71] filter image-text pairs from PMC-15M [126], and train a biomedical-specialized LVLM with a small amount data based on the LLaVA-pretrained parameter. Zhang *et al.* generate a large-scale medical VQA dataset, PMC-VQA [127], through the self-instruction on PMC-OA [74]. Leveraging PMC-VQA, Zhang *et al.* train a biomedical-specialized VQA model, termed MedVInT, which achieves state-of-the-art performance on many Medical VQA datasets. Furthermore, they continue to propose RadFM [117], the first multi-modal foundation model for seamlessly integrating natural languages with both 2D and 3D radiologic images, which better fits the medical practical. Overall, we compare the information of aforementioned LVLMs in Table 2.

Despite the increasing attention, it still lacks a comprehensive evaluation of LVLMs within the medical domain. To address this deficiency, we perform a thorough evaluation for these LVLMs in this paper.

2.2. Medical VQA Dataset

With the rapid development of LVLMs, the field of Medical VQA has received considerable interest in recent years. VQA-RAD [69], SLAKE [75], Path-VQA [122], and VQA-Med [24] are four widely used Medical VQA datasets. However, they all have less than 5K images. Meanwhile, VQA-RAD and SLAKE only contain images captured by CT, MRI or X-Ray, reducing their diversity and restricting their further application. Moreover, all these datasets only cover limited parts of the human anatomy, hindering a comprehensive evaluation across various human anatomical regions. For example, VQA-RAD only contains images of the head, chest, and abdomen, while Slake only covers the head, chest, abdomen, pelvis, and neck, lacking images of other important parts of the human body. Recently, motivated by the success of self-instruction in textual data generating, PMC-VQA, a large-scale Medical VQA dataset, is proposed based on numerous image-caption pairs. However, its images and text are extracted from online papers, which can result in image compression and a significant gap from real-world medical applications.

In this paper, we collect a new large-scale OmniMed-VQA dataset, which contains medical images captured through 12 different modalities and covers almost every anatomical region of the human body. OmniMedVQA is the current largest MedVQA dataset with real medical images. We hope it can help the community better evaluate the fundamental ability of LVLMs in the biomedical field.

3. Dataset Collection

In this part, we introduce the collection process of our OmniMedVQA dataset. To make full use of real medical images, we collect an enormous medical classification dataset and construct the question-answer pairs based on their inherent attributes using the ChatGPT API. Generally speaking, the construction process has the following four steps.

- **Original dataset preparing.** Due to the well-known difficulties in downloading medical datasets, it is notably time-consuming to obtain plentiful and suitable datasets. To construct a comprehensive VQA benchmark, we collected 73 different medical classification datasets encompassing 12 different imaging modalities, which span more

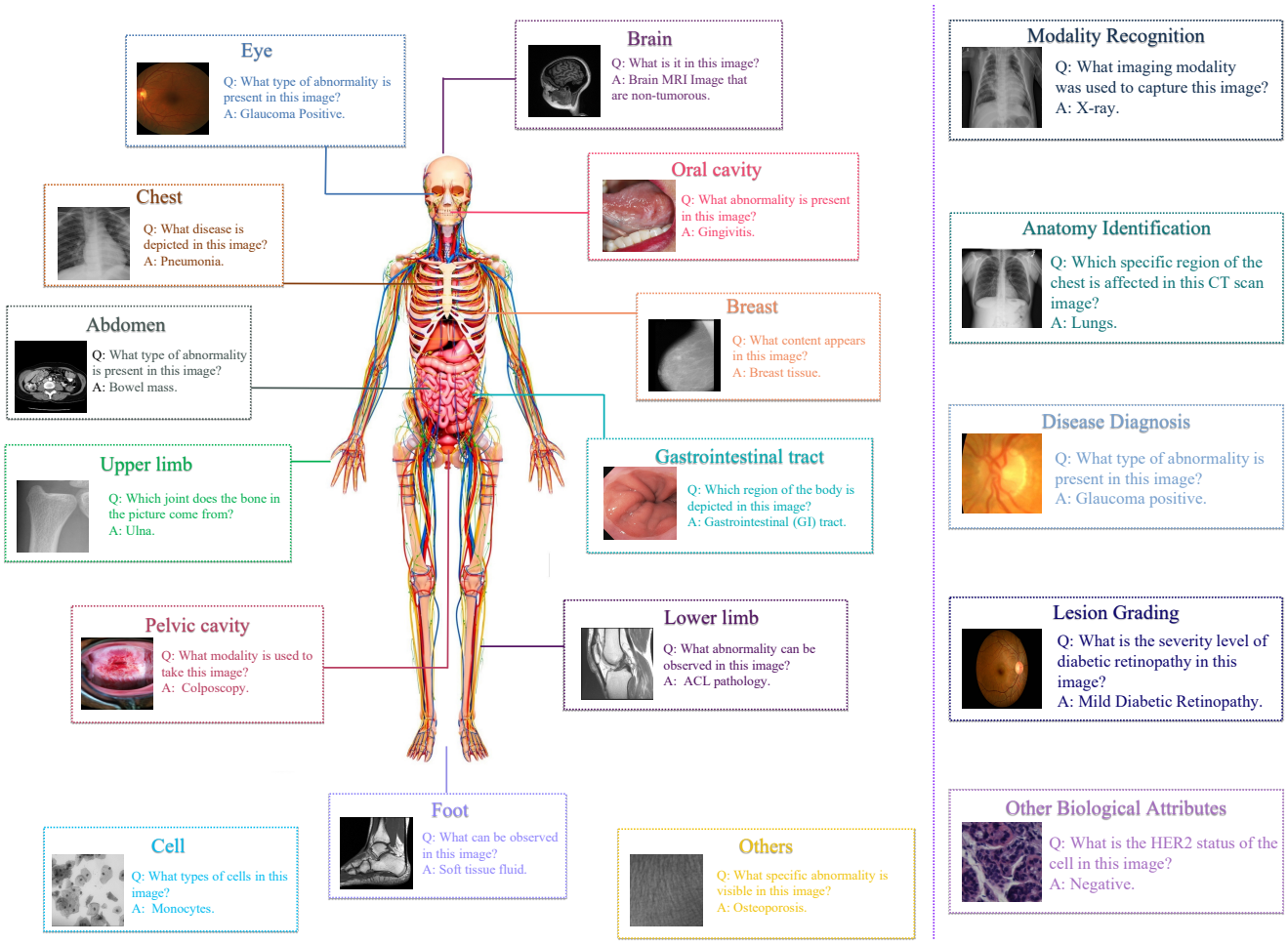

Figure 1. **Left:** Overview of our OmniMedVQA dataset. OmniMedVQA covers the majority of radiologic modalities and anatomical regions of the human body, such as the brain, eyes, oral cavity, chest, breast, abdomen, upper limb, lower limb, feet, etc. **Right:** Illustrations of samples from five different question types.

than 20 different human anatomical regions. The details of all the involved datasets are presented in the supplementary material.

- **Design QA templates.** Based on the collected datasets, we need to transfer the original classification attributes into QA format. To achieve this goal, we first construct the QA template for each dataset. On the one hand, the category information is naturally suitable for constructing a QA-pair. Therefore, we design question templates according to their original categories. For example, in the MAlig_Lymph dataset[89], it contains 3 different diseases. Therefore, we could design a QA template as "Q: What is the specific diagnosis for the cancer cells in this image?; A: Chronic Lymphocytic Leukemia.". On the other hand, through further understanding of the dataset, we construct QA pairs according to their other attributes, such as modality and anatomy information. For example, in SARS-COV-2 Ct-Scan dataset[107], we could also ask "What is the modality of the image?" or "What is the abnormal organ in the picture?", which evaluate the ability of modality recognition and anatomy localization. In summary, all QA pairs fall into five distinct question types: Modality Recognition, Anatomy Identification, Disease Diagnosis, Lesion grading and Other Biological Attributes. We illustrate the sample of each question type in Fig 1. As depicted in the right part of Fig 1, each question type could evaluate one specific capability within the biomedical field. Specifically, Lesion Grading aims to assess the severity of lesions in the images, while Other Biological Attributes include the analysis of various attributes related to medical images, such as cell shape, cancer status, imaging direction *et al.* The detailed statistic information of each type is listed in Table 3. Furthermore, we meticulously control the number of different items from each template to ensure balance and prevent significant bias. Specifically, we construct

Table 3. The number of images and items for different question types in OmniMedVQA.

Question Type	# Images	# Items
Modality Recognition	19,381	19,427
Anatomy Identification	19,992	20,330
Disease Diagnosis	73,099	73,455
Lesion Grading	2621	2621
Other Biological Attributes	12,156	12,162
Total	118,010	127,995

QA templates from 73 medical classification datasets and calculate the number of potential images that could be used under each template. Then, we select the images using the Inverse Proportional Sampling strategy. Namely, templates with a larger number of associated images are assigned a smaller sample ratio. In this way, our dataset keeps a balanced distribution across categories and avoids the bias on some reduplicate QAs, ensuring OmniMedVQA as a diverse and comprehensive dataset.

- **Refine QA pairs.** In order to increase the diversity of our dataset and better evaluate the capability of each LVLM, we employ the ChatGPT-3.5 API to perform two key operations. First, we reformulate the question in each item to change the expression style and syntactic structure, while preserving the original semantic meaning, which allows us to evaluate the adaptability of LVLMs to various linguistic representations. Second, we leverage the GPT-3.5 API to generate a set of incorrect options for each item, which are utilized to construct multiple-choice question-answer pairs. Specifically, each item in our dataset is paired with incorrect answers as the candidate options. The number of options varies from 2 to 4, which depends on the content of the specific question. By doing so, it is more convenient to judge the correctness of the response from LVLM. The generation process of each item are depicted in Fig. 2.
- **Human double check.** To ensure data quality, we conducted further inspections to guarantee the validity of our OmniMedVQA dataset.

Overall, OmniMedVQA contains 118,010 images with 127,995 QA-items, covering 12 different modalities and referring to more than 20 human anatomical regions. The detailed modality and anatomy information of our dataset are listed in Table 4. We want to emphasize that while the classification attributes may appear intuitive and not overly complex, they play a crucial role in evaluating the fundamental capabilities of LVLMs in the medical domain, which are critical to support broader applications in this field. Meanwhile, the evaluation results reveal that the medical-specialized LVLM cannot handle these questions well, indicating a deficiency in its foundational medical knowledge and highlighting the need for more versatile LVLMs.

Table 4. The modalities and anatomies involved in our OmniMedVQA dataset.

Modality	Colposcopy, CT (Computed Tomography), Digital Photography, Fundus Photography, Infrared Reflectance Imaging, MR (Magnetic Resonance Imaging), OCT (Optical Coherence Tomography), Dermoscopy, Endoscopy, Microscopy Images, X-Ray, Ultrasound
Anatomy	Lung, Mammary Gland, Hand, Upper Limb, Eye, Uterus, Intestine, Skin, Shoulder, Kidney, Gallbladder, Pancreas, Spleen, Liver, Pelvic, Ovary, Blood Vessel, Spine, Urinary System, Adipose Tissue, Muscle Tissue, Oral Cavity, Knee, Foot, Lower Limb

4. Evaluation Method

As mentioned in Sec. 3, for the purpose of evaluation convenience, we provide each QA pair with incorrect answers as candidate options, resulting in a multi-choice Question-Answer task. However, some LVLMs, especially the medical-specialized LVLMs, exhibit poor instruction-following performance during the evaluation, failing to generate responses according to the given options. We think that this situation does not necessarily imply that these models lack medical knowledge, which may simply indicate that they are not proficient in processing input in the form of multiple-choice questions. Therefore, to ensure a more fair comparison, we adopt two different metrics in our evaluation, Question-answering Score and Prefix-based Score [73, 120]. Their evaluation processes are depicted in Fig. 3, and we report their performances respectively in Sec. 5.

4.1. Question-answering Score

Given an input image with the question expressions and candidate options, we first combine them to construct the prompt for LVLM. For example, we can utilize the prompt template: "This is a medical question with several Options, and there is only one correct answer among these options. Please select the correct answer for the question. Remember, you can only select one option. The Question is:<Question>. ### The candidate Options are:<Options>", based on this template, we insert the current question and candidate options. Then, we deliver the image with the prompt to the LVLM to generate the response. Afterwards, following previous works [120], we calculate the similarity of the response with the candidate options and select the option with the largest similarity as the final prediction. Finally, we compare the prediction with the ground-truth answer and judge the correctness.

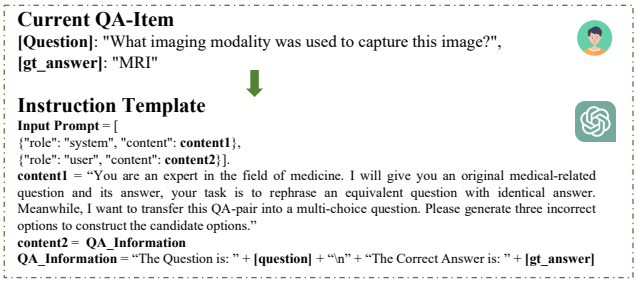

Figure 2. The illustration of the process by which we transfer the QA item from the QA template into the multi-choice question-answer pair.

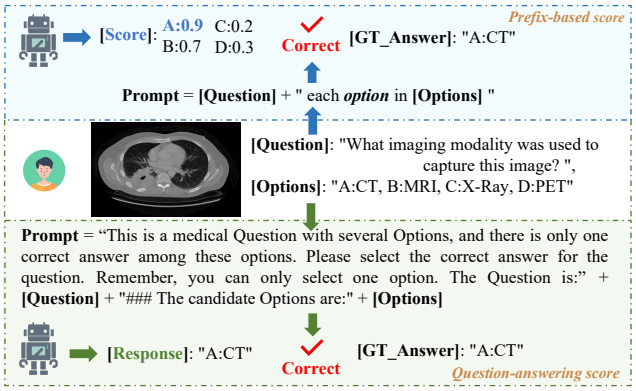

Figure 3. Evaluation process when adopting Question-answering score and Prefix-based score respectively.

4.2. Prefix-based Score

We also utilize the prefix-based score to evaluate the inherent biomedical knowledge and avoid hallucination in the response. Specifically, in the context of multi-choice Question-Answer tasks, given an input image with the textual sentence, we first extract the visual features and text embeddings, respectively. Then, the visual features are prefixed into the text embeddings, which are subsequently delivered into the LLM to calculate the likelihood score [120]. This score is considered the prefix-based score for this image-text pair, which reflects the probability of the model generating the corresponding textual content. Therefore, for each candidate option within the specific item, we combine it with the question and then calculate the prefix-based score. The option yielding the highest prefix-based score means it is the most likely answer for this question, which is considered as the final answer of the corresponding LVLM. Then, we compare the final answer with the ground-truth answer to judge the correctness, based on which we compute the VQA accuracy.

It is worthwhile to mention that, although the prefix-based score is not directly equivalent to the response from the LVLM, it measures the likelihood of each option being regarded as the correct answer, which reflects the level of inherent knowledge of the model. In fact, during our evaluation, our primary objective is to find the shortcomings of existing LVLMs in the biomedical field and propose insightful suggestions for future research. However, considering the challenges of OmniMedVQA, it is really hard for these LVLMs to directly generate the correct answer for all the questions. Therefore, the prefix-based score aligns with our evaluation criteria and is a fair metric for the evaluation.

5. Experiment

5.1. Experimental Details

In this section, we perform zero-shot evaluation to assess 12 representative LVLMs. All the experimental environments and hyper-parameters are set according to their released code. Specifically, since MedVInT [127] has two variants versions that exhibit different capabilities in open-ended and multiple-choice tasks tasks, we employ MedVInT-TD and MedVInT-TE for the evaluation via Question-answering Score and Prefix-based Score, respectively.

5.2. Overall Performance

The evaluation results are listed in Table 5 which is divided into two sections based on rows. The first eight rows present the accuracy of different general-purpose LVLMs, while the last four rows reflect the performance of medical-domain LVLMs. Specifically, we report the Question-answering Score and Prefix-based Score separately in Table 5. Generally speaking, our OmniMedVQA is extremely challenging, with most LVLMs only slightly surpassing the performance of random guess. Moreover, we have several observations based on the results in Table 5.

1. To our surprise, the general-domain LVLM, BLIP2 [73], achieves the best performance for all tasks on average, surpassing all tested LVLM models in the medical domain by a large margin. This suggests that emerging property does not occur when using the current medical data to adapt general-purpose LVLMs for medical tasks.
2. A strong model in aligning image-text pairs in the medical domain is urgently needed for medical-domain LVLMs. BLIP performs well in both general-purpose QA dataset [104, 120] and our OmniMedQA because it is trained by massive high-quality image-text pairs in various visual domains. Hence, the key to developing general medical-purpose LVLMs lies in training models with massive high-quality image captioning data from various medical domains.
3. MedVInT and Med-Flamingo achieve the highest overall accuracy among all evaluated medical LVLMs and also outperforms many general-purpose LVLMs except for BLIP2 and InstructBLIP. This success may be at-

Table 5. The accuracy of representative LVLMs on our OmniMedVQA in terms of five different question types. Notably, we report the Question-answering Score and Prefix-based Score before and after "/", respectively. Meanwhile, in each column, the best performance is marked in red, while the second best performance is marked in blue.

Model	Modality Recognition	Anatomy Identification	Disease Diagnosis	Lesion Grading	Other Biological Attributes	Overall
Random Guess	25.00	25.84	28.41	25.40	37.49	28.28
MiniGPT-4 [128]	28.23 / 28.71	28.41 / 30.26	30.51 / 20.24	32.85 / 38.53	37.30 / 43.28	30.53 / 25.68
BLIP-2 [73]	57.51 / 37.85	49.19 / 64.75	46.24 / 23.19	30.52 / 25.03	73.52 / 37.70	50.69 / 33.43
InstructBLIP [41]	70.62 / 24.83	42.75 / 51.78	33.62 / 22.41	54.60 / 54.71	48.16 / 28.85	42.49 / 28.71
mPLUG-Owl [123]	27.93 / 16.46	24.44 / 30.45	30.31 / 29.32	38.50 / 76.96	37.84 / 43.70	29.90 / 29.89
Otter [70]	25.62 / 10.53	25.14 / 23.47	27.12 / 22.11	34.22 / 39.53	32.12 / 24.74	27.20 / 21.17
LLaVA [77]	27.21 / 14.43	25.99 / 15.08	27.35 / 19.60	38.53 / 26.82	32.68 / 38.24	27.85 / 20.02
LLaMA_Adapter_v2 [48]	44.51 / 33.48	33.73 / 40.67	29.17 / 24.93	39.07 / 38.65	36.80 / 34.07	33.15 / 29.88
VPGTrans [125]	29.32 / 31.77	30.76 / 36.27	26.91 / 18.88	29.61 / 38.53	32.65 / 41.47	28.49 / 26.15
Med-Flamingo [84]	28.67 / 21.21	25.32 / 25.16	41.47 / 26.10	31.25 / 49.60	35.27 / 34.97	36.17 / 26.54
RadFM [117]	21.31 / 38.57	19.96 / 29.48	28.46 / 25.66	24.72 / 35.86	37.58 / 31.73	26.82 / 29.00
MedVInT [127]	59.79 / 30.92	41.36 / 23.90	36.79 / 25.02	15.49 / 5.23	46.79 / 30.12	41.50 / 25.81
LLaVA-Med [71]	31.38 / 12.02	28.34 / 22.06	28.01 / 27.25	32.35 / 30.98	29.23 / 25.88	28.78 / 24.06

Table 6. The overall accuracy of representative LVLMs on our OmniMedVQA in terms of different modalities. Here, we report the accuracy of all items within each modality when utilizing the **Question-answering score**. Specifically, **Co** denotes Colposcopy, **CT** denotes Computed Tomography, **DP** denotes Digital Photography, **FP** denotes Fundus Photography, **IRI** denotes Infrared Reflectance Imaging, **MR** denotes Magnetic Resonance Imaging, **OCT** denotes Optical Coherence Tomography, **Der** denotes Dermoscopy, **End** denotes Endoscopy, **Mic** denotes Microscopy Images, **US** denotes Ultrasound. Meanwhile, in each column, the best and second-best performance are marked in red and blue, respectively.

Model	Co	CT	DP	FP	IRI	MR	OCT	Der	End	Mic	X-Ray	US
MiniGPT-4 [128]	23.67	22.81	18.05	42.37	38.51	27.60	31.40	40.09	30.26	28.05	39.75	25.50
BLIP-2 [73]	48.52	56.74	23.01	57.66	66.18	41.77	68.08	41.07	48.85	50.17	70.55	37.27
InstructBLIP [41]	32.25	28.72	35.75	37.72	59.27	33.79	42.59	61.86	36.65	48.20	61.21	41.25
mPLUG-Owl [123]	36.69	24.54	19.81	41.81	38.44	29.82	43.76	35.98	24.45	25.99	28.29	21.40
Otter [70]	33.73	18.53	18.20	37.70	30.70	26.37	29.64	42.66	33.94	22.94	31.73	23.49
LLaVA [77]	12.72	17.73	22.25	32.16	31.23	26.99	33.73	49.67	38.20	27.95	31.35	18.66
LLaMA_Adapter_v2 [48]	38.46	21.41	28.93	36.85	35.70	27.23	33.00	51.43	46.62	34.78	46.70	34.05
VPGTrans [125]	32.54	21.26	20.10	34.50	32.60	25.36	25.14	44.66	30.53	23.61	46.53	25.45
Med-Flamingo [84]	18.40	38.47	21.48	27.61	39.69	40.01	26.51	32.33	30.97	46.60	28.30	24.64
RadFM [117]	15.43	27.44	13.25	28.99	36.13	24.16	32.80	39.03	28.40	24.81	29.21	16.57
MedVInT [127]	39.17	40.74	43.89	39.69	46.22	42.84	23.26	29.13	30.11	40.71	56.62	41.26
LLaVA-Med [71]	28.99	18.69	18.34	35.14	30.68	27.49	34.61	44.90	41.88	26.33	31.26	29.88

tributed to the extensive medical knowledge they are injected in the training. Med-Flamingo learns from more than 4k textbooks while MedVInT is trained based on 381K image-caption pairs. This indicates that, to obtain a better performance, more knowledge in the medical domain should be injected into the LVLMs.

4. Through a comparative analysis of LLaVA and LLaVA-Med, we conclude that medical instruction tuning can improve the performance of general-purpose LVLM in the biomedical field. However, when comparing different medical LVLMs, we can find that LLaVA-Med delivers the worst performance. As listed in Table 2, LLaVA-Med initiates its model from the pre-trained LLaVA [71] and incorporates only a small amount of data in the training, through which they expect to save the training cost. However, the evaluation results suggest the LLaVA pre-trained model is not suitable and the advanced performance should be instructed by sufficient data, which calls for a robust pre-trained model tailored to the biomedical field and high-quality data. In fact, the success of MiniGPT-4 in the general domain demonstrates that high-quality data, even if it is in small quantities, could strongly support the training of LVLM. However, the instruction data for LLaVA-Med is generated through the GPT-4 API and only relies on textual information. We believe this substantially compromises

Table 7. The overall accuracy of representative LVLMs on our OmniMedVQA in terms of different modalities. Here, we report the accuracy of all items within each modality when utilizing **Prefix-based score**. Specifically, **Co** denotes Colposcopy, **CT** denotes Computed Tomography, **DP** denotes Digital Photography, **FP** denotes Fundus Photography, **IRI** denotes Infrared Reflectance Imaging, **MR** denotes Magnetic Resonance Imaging, **OCT** denotes Optical Coherence Tomography, **Der** denotes Dermoscopy, **End** denotes Endoscopy, **Mic** denotes Microscopy Images, **US** denotes Ultrasound. Meanwhile, in each column, the best and second-best performance are marked in red and blue, respectively.

Model	Co	CT	DP	FP	IRI	MR	OCT	Der	End	Mic	X-Ray	US
MiniGPT-4 [128]	26.33	29.46	22.94	23.59	43.61	12.75	30.00	25.18	26.90	27.60	40.11	27.20
BLIP-2 [73]	32.25	38.87	25.95	20.01	43.40	20.45	18.70	19.80	25.95	28.43	49.81	81.79
InstructBLIP [41]	17.46	35.66	10.01	26.74	27.26	15.02	51.74	29.53	28.77	22.80	40.85	58.51
mPLUG-Owl [123]	18.64	37.00	18.31	56.87	44.96	13.22	41.76	18.75	35.32	31.04	32.18	29.14
Otter [70]	2.37	32.55	20.78	16.96	21.90	10.73	45.98	22.52	23.39	21.99	27.80	20.88
LLaVA [77]	33.73	38.27	23.55	13.84	36.37	4.65	51.23	14.17	22.16	17.65	24.10	20.74
LLaMA_Adapter_v2 [48]	28.40	36.03	21.36	22.11	33.69	17.26	54.22	21.61	32.23	30.60	41.71	47.51
VPGTrans [125]	35.21	30.30	27.71	23.93	42.35	11.04	31.60	22.18	24.77	25.87	39.94	40.89
Med-Flamingo [84]	30.56	22.25	7.40	51.57	30.22	14.43	58.57	39.48	46.61	23.04	38.13	17.42
RadFM [117]	22.55	45.47	13.43	21.48	25.16	24.52	37.40	25.71	34.44	20.52	55.21	24.78
MedVInT [127]	60.24	37.77	39.04	9.45	32.92	30.10	18.54	20.72	24.03	15.35	29.09	25.43
LLaVA-Med [71]	13.61	36.75	10.48	16.15	21.61	24.58	51.51	25.35	27.92	17.40	25.54	16.75

data quality, leading to the diminished accuracy. This underscores the need for an effective caption-generation model like BLIP2 [73], which can generate accurate and detailed textual descriptions for biomedical images.

Overall, through our evaluation, we find that medical-specialized LVLMs do not present outstanding performance. For most question types, two general LVLMs, BLIP2 and InstructBLIP obtain the best accuracy. To better elaborate the underlying reasons, we conduct an in-depth analysis of the evaluation results in Sec. 5.3.

5.3. Analysis in terms of modalities

To further analyze the performance of LVLMs in the biomedical field, we report the accuracy of all QA items in terms of different modalities in Table 6 and Table 7, which present the Question-answering score and Prefix-based score respectively. Based on the results, we have following two observations.

1. Although medical LVLMs exhibit lower accuracy when considering the overall dataset, they tend to perform well in modalities characterized by substantial differences from general images, such as CT and MRI. However, in modalities with similar distributions to those in general domain images, medical-specialized LVLMs fail to demonstrate notably superior performance.
2. RadFM aims to initiate the development of the radiology foundation model, which is trained with more than 19M radiologic image-test pairs. Among their dataset, CT, MRI and X-Ray constitute a significant proportion. Therefore, as listed in Table 7, RadFM achieves the best performance on CT and X-Ray tasks, and obtain competitive performance on MR task. This reveals the potential for performance improvement with high-quality instruction data. To bring an all-around LVLM in the biomedical field, the inclusion of additional high-quality data from various modalities, such as Fundus Photography and Infrared Reflectance Imaging, is imperative.

6. Conclusion

Since LVLMs have recently received remarkable attention in the whole community, this paper aims to evaluate their performance in the biomedical field. To achieve this goal, we collect OmniMedVQA, a large-scale medical VQA dataset. OmniMedVQA has 118,010 images with 127,995 question-answer items, which include 12 different modalities and cover more than 20 human anatomical regions. Therefore, OmniMedVQA could support the throughout evaluation of different LVLMs. During the evaluation, we assess 12 different LVLMs, including 8 general-domain models and 4 specialized LVLMs for the biomedical field. To our great surprise, despite their claims of robustness, the medical LVLMs exhibit inferior performance to those general-domain models, which reveals the shortcomings of these medical models. We point out that to become a more versatile medical expert, medical LVLMs consistently require additional knowledge of specific modalities, such as Infrared Reflectance Imaging and Fundus Photography, on which the performance of medical LVLMs is significantly inferior to general-domain models. We hope our dataset provides a comprehensive evaluation benchmark for medical LVLMs and our findings offer useful suggestions for future research.

Acknowledgement

This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

References

[1] Chest ct-scan images dataset. https://tianchi.aliyun.com/dataset/93929,. 2

[2] Covid ct dataset. https://tianchi.aliyun.com/dataset/106604,. 2

[3] Isic 2019 challenge. https://challenge.isic-archive.com/landing/2019/. 2

[4] Oral cancer (lips and tongue) images. https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images,. 2

[5] Dental condition dataset. https://www.kaggle.com/datasets/salmansajid05/oral-diseases,. 2

[6] Glaucoma grading based on multi-modality images. https://aistudio.baidu.com/competition/detail/119/0/task-definition,. 2

[7] Glaucoma detection. https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection,. 2

[8] Analysis of images to detect abnormalities in endoscopy. https://aidasub-cleceliachy.grand-challenge.org/Description/, 2016. 2

[9] Diabetic retinopathy arranged - retina images with class labels for classification. https://tianchi.aliyun.com/dataset/93926, 2023. 2

[10] Cataract image dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset, 2023. 2

[11] Bright challenge: Breast tumor image classification on gigapixel histopathological images. https://research.ibm.com/haifa/Workshops/BRIGHT/, 2023. 2

[12] Nlm - malaria data. https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html, 2023. 2

[13] Blood cell images. https://www.kaggle.com/datasets/paultimothymooney/blood-cells, 2023. 2

[14] Retinal oct - c8 dataset. https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8/data, 2023. 2

[15] X-ray hand small joint classification dataset (based on bone age scoring method rus-chn). https://aistudio.baidu.com/datasetdetail/69582/0, 2023. 2

[16] Covid-19 image dataset: 3 way classification - covid-19, viral pneumonia, normal. https://tianchi.aliyun.com/dataset/93853, 2023. 3

[17] Michael D Abràmoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131 (3):351–357, 2013. 2

[18] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, 2022. 3

[19] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 1

[20] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019. 2

[21] Amanullah Asraf and Zabirul Islam. Covid19, pneumonia and normal chest x-ray pa dataset. *Mendeley Data*, 1, 2021. 3

[22] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):1–13, 2017. 1

[23] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 1

[24] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021. 1, 3

[25] BenO, JL Jones, H Kumar, Meg Risdal, M Rao, Vadim Sherman, Vipul, Wendy Kan, and Yau Ben-Or. Intel & mobileodt cervical cancer screening. https://kaggle.com/competitions/intel-mobileodt-cervical-cancer-screening, 2017. 2

[26] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. 2

[27] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[28] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017. 1

[29] Kimberlin van Wijnen Carole Sudre. Where is valdo -

vascular lesions detection challenge 2021. https://valdo.grand-challenge.org/, 2021. 1

[30] Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, 12(1):4828, 2021. 2

[31] Carlo W Cereda, Søren Christensen, Bruce CV Campbell, Nishant K Mishra, Michael Mlynash, Christopher Levi, Matus Straka, Max Wintermark, Roland Bammer, Gregory W Albers, et al. A benchmarking tool to evaluate computer tomography perfusion infarct core predictions against a dwi standard. *Journal of Cerebral Blood Flow & Metabolism*, 36(10):1780–1789, 2016. 1

[32] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2

[33] Pingjun Chen. Knee osteoarthritis severity grading dataset. *Mendeley Data*, 1:21–23, 2018. 2

[34] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[35] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 3

[36] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 2

[37] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020. 3

[38] Olivier Commowick, Michaël Kain, Romain Casey, Roxana Ameli, Jean-Christophe Ferré, Anne Kerbrat, Thomas Tourdias, Frédéric Cervenansky, Sorina Camarasu-Pop, Tristan Glatard, et al. Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset. *Neuroimage*, 244:118589, 2021. 1

[39] Will Cukierski. Histopathologic cancer detection. https://kaggle.com/competitions/histopathologic-cancer-detection, 2018. 2

[40] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022. 2

[41] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 3, 7, 8, 6

[42] Coen De Vente, Koenraad A Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, et al. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 2023. 2

[43] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014. 2

[44] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, 18: 1–19, 2019. 2

[45] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenole Quellec, Sarah Matta, et al. Adam challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022. 2

[46] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Palm: Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *arXiv preprint arXiv:2305.07816*, 2023. 2

[47] Radovan Fusek. Pupil localization using geodesic distance. In *Advances in Visual Computing: 13th International Symposium, ISVC 2018, Las Vegas, NV, USA, November 19–21, 2018, Proceedings 13*, pages 433–444. Springer, 2018. 2

[48] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2, 3, 7, 8, 6

[49] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. Mednode: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42(19):6578–6585, 2015. 2

[50] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021. 2

[51] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942*, 2022. 2

[52] Anubha Gupta and Ritu Gupta. Isbi 2019 c-nmc challenge: Classification in cancer cell imaging. *Select Proceedings*, 2019. 2

[53] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2

[54] Arsany Hakim, Søren Christensen, Stefan Winzeck, Maarten G Lansberg, Mark W Parsons, Christian Lucas, David Robben, Roland Wiest, Mauricio Reyes, and Greg Zaharchuk. Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the isles challenge. *Stroke*, 52(7):2328–2337, 2021. 1

[55] Ahmed Hamada. Br35h:: Brain tumor detection 2020. *Version 5*, 2020. 2

[56] Khaled Harrar. Texture characterization of bone radiograph images. application to osteoporosis diagnosis. 2014. 2

[57] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 1

[58] Towhidul Islam, Mohammad Arafat Hussain, Forhad Uddin Hasan Chowdhury, and BM Riazul Islam. A web-scraped skin image database of monkeypox, chickenpox, smallpox, cowpox, and measles. *biorxiv*, pages 2022–08, 2022. 2

[59] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4 (6):475, 2014. 3

[60] Soroush Javadi and Seyed Abolghasem Mirroshandel. A novel deep learning method for automatic assessment of human sperm images. *Computers in biology and medicine*, 109:182–194, 2019. 2

[61] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 36722–36732, 2022. 1

[62] Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. Kaggle, 2019. 2

[63] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018. 2

[64] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018. 2

[65] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, and Philippe Büchler. The virtual skeleton database: an open access repository for biomedical research and collaboration. *Journal of medical Internet research*, 15 (11):e245, 2013. 1

[66] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[67] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2023. 1

[68] Hugo J Kuijf and E Bennink. Grand challenge on mr brain segmentation at miccai 2018, 2019. 1

[69] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 1, 3

[70] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2, 3, 7, 8, 6

[71] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 1, 2, 3, 7, 8, 6

[72] Fei Li, Diping Song, Han Chen, Jian Xiong, Xingyi Li, Hua Zhong, Guangxian Tang, Sujie Fan, Dennis SC Lam, Weihua Pan, et al. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. *NPJ digital medicine*, 3(1):123, 2020. 2

[73] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5, 6, 7, 8

[74] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023. 3

[75] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 1, 3

[76] Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *Plos one*, 14(9):e0222025, 2019. 2

[77] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 7, 8, 6

[78] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6), 2022. 2

[79] Oskar Maier, Bjoern H Menze, Janina Von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017. 1

[80] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315, 2022. 1, 2

[81] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph^2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013. 2

[82] Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience*, 2015:1–1, 2015. 1

[83] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 1

[84] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. 1, 2, 3, 7, 8, 6

[85] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[86] Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS One*, 11(2): e0149399, 2016. 2

[87] Thivya Narendran. Image set for retinal artery-vein nicking assessment. https://people.eng.unimelb.edu.au/thivun/projects/AV_nicking_quantification/. 2

[88] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 2

[89] Nikita V Orlov, Wayne W Chen, David Mark Eckley, Tomasz J Macura, Lior Shamir, Elaine S Jaffe, and Ilya G Goldberg. Automatic classification of lymphoma images with transform-based global features. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):1003–1013, 2010. 4, 2

[90] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 3

[91] Andre GC Pacheco and Renato A Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in biology and medicine*, 116:103545, 2020. 2

[92] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2

[93] Hady Ahmady Phoulady and Peter R. Mouton. A new cervical cytology dataset for nucleus detection and image classification (cervix93) and methods for cervical nucleus detection, 2018. 2

[94] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017. 2

[95] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022. 2

[96] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017. 3

[97] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021. 2

[98] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis*, 73: 102155, 2021. 1

[99] Fabio Scarpa, Enrico Grisan, and Alfredo Ruggeri. Automatic recognition of corneal nerve structures in images

from confocal microscopy. *Investigative ophthalmology & visual science*, 49(11):4801–4807, 2008. 2

[100] Fabio Scarpa, Xiaodong Zheng, Yuichi Ohashi, and Alfredo Ruggeri. Automatic evaluation of corneal nerve tortuosity in images from in vivo confocal microscopy. *Investigative ophthalmology & visual science*, 52(9):6404–6408, 2011. 2

[101] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[102] Uğur Şevik, Cemal Köse, Tolga Berber, and Hidayet Erdöl. Identification of suitable fundus images using automated quality assessment methods. *Journal of biomedical optics*, 19(4):046006–046006, 2014. 2

[103] Fariba Shaker, S Amirhassan Monadjemi, Javad Alirezaie, and Ahmad Reza Naghsh-Nilchi. A dictionary learning approach for human sperm heads classification. *Computers in biology and medicine*, 91:181–190, 2017. 2

[104] Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023. 6

[105] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[106] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000. 3

[107] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pages 2020–04, 2020. 4, 2

[108] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015. 2

[109] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, pages 375–378, 1994. 3

[110] Siham Tabik, Anabel Gómez-Ríos, José Luis Martín-Rodríguez, Iván Sevillano-García, Manuel Rey-Area, David Charte, Emilio Guirado, Juan-Luis Suárez, Julián Luengo, MA Valero-González, et al. Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images. *IEEE journal of biomedical and health informatics*, 24(12):3595–3605, 2020. 3

[111] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[112] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 2

[113] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020. 3

[114] Shuo Wang, Chen Qin, Chengyan Wang, Kang Wang, Haoran Wang, Chen Chen, Cheng Ouyang, Xutong Kuang, Chengliang Dai, Yuanhan Mo, et al. The extreme cardiac mri analysis challenge under respiratory motion (cmrxmotion). *arXiv preprint arXiv:2210.06385*, 2022. 1

[115] Stefan Winzeck, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, et al. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri. *Frontiers in neurology*, 9:679, 2018. 1

[116] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023. 1

[117] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 1, 2, 3, 7, 8, 6

[118] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021. 1

[119] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 11:759007, 2021. 2

[120] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 2, 5, 6

[121] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 1

[122] He Xuehai, Zhang Yichen, Mou Luntian, Xing Eric, and Xie Pengtao. Pathvqa: 30000+ questions for medical vi-

sual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1, 3

[123] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 3, 7, 8, 6

[124] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation, 2019. 3

[125] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023. 2, 3, 7, 8, 6

[126] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 3

[127] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 1, 2, 3, 6, 7, 8

[128] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 3, 7, 8, 6